Transcript for Episode 85:

Powering the Future of AI

With Jane Ross, Tim Horan, Colin Rusch, and Noah Kaye

Jane Ross:

Hello everybody, and welcome back to Let's Talk Future. I'm your host, Jane Ross, and the issue we'll attack here is the promise of generative AI versus the reality of how to make it all happen. We've already done quite a few episodes about the potential opportunities created by generative AI, but the devil is in the details. How are we going to create enough data centers and fuel them responsibly and reliably? How will we provide backup and cooling, and the systems to deploy AI for all of the applications that we're bullish about? Well, here we have a top-notch interdisciplinary team from Oppenheimer to address these issues, Tim Horan, Colin Rush and Noah Kaye, representing cloud communications and sustainable growth research silos at Oppenheimer. So welcome back, Tim Colin and Noah, and before we get into the particulars of AI and energy infrastructure, I'm going to ask each of you to highlight the companies that you are recommending investors pay attention to in all of this. So Tim, starting with you and the cloud and communications realm.

Tim Horan:

So three top picks here. Microsoft really is the dominant AI infrastructure company. One of their primary suppliers is Equinix, that basically runs all the neutral data centers to enable all the interconnection of traffic that's going to be required by AI. And then the third company would be CloudFlare. CloudFlare is building out edge compute infrastructure, and we're going to have a lot of applications that are going to be global, that are going to require lower latency, and you're going to want to have a consistent developer platform, which is what CloudFlare does. That's it, Jane.

Jane Ross:

Perfect. Okay. Noah, over to you and the infrastructure plays.

Noah Kaye:

Thanks, Jane. The first one's going to be Vertiv, ticker, VRT. A pretty well-known name at this point critical digital infrastructure for data centers, communication networks and CNI environments. And Vertiv, 75% of their exposure in data centers, their AI-related pipeline has doubled in the last quarter. The second one would be Modine, ticker, MOD. Modine is a leading provider of thermal management services for both the mobility, HVAC and increasingly, data center markets. Data centers are about 11% of their sales. They're growing 70% per year. And then the third one actually outside of the traditional equipment suppliers would be Itron, which is a leading smart grid player, and we think a key orchestrator of balancing supply and demand as we look at some of these challenges facing the grid.

Jane Ross:

Got it. And geez, pipeline doubled in the last quarter. This is what we're talking about. Okay. Colin, what about you and the renewable space?

Colin Rusch:

I'll start with the grid-related plays on capacity with American Superconductor enabling higher power quality and increased grid capacity with the right solution. And then generation with First Solar as a solar play, and Canadian Solar, CSIQ, as one of the largest developers of both solar and energy storage products.

On the control side, dovetailing off of what Noah is saying on Itron at the edge of the grid, we're looking at Enphase as a leading controls play relative to solar and energy storage integration at the residential level. Sunrun as a portfolio play on relieving stress from the network by implementing virtual power plants, and then on the mobile side, as we see distributed inference and incremental compute out at the device level, Equinix, ENVX.

Jane Ross:

Okay. Well, great. Thank you. I just love getting those names out there up front so investors know who to think about. Okay, so let's start with you, Tim, and step back. Your team has done a ton of work to create estimates for AI revenue, for chip forecasts, and you've talked about the imbalance between the data center power supply and demand landscape. So if you could orient us about all of that and what's setting up all of this demand?

Tim Horan:

Totally right. So the promise of AI is that it could add about a trillion dollars a year, give or take, to global economic growth, and it can enable basically 10 to 30% improvement in productivity per employee globally over a couple year period. That's going to require an awful lot of compute infrastructure. We estimate the amount of spending on compute, and that would include data centers, for generative AI was about $60 billion last year. In 2027, we think that'll increase fourfold to about $250 billion, which is basically a once in a century type capital spend on a once a century kind of killer application, which AI basically is, and we derive a lot of that from just the GPU forecasts of what the dominant computer chip manufacturers are basically telling their vendors and their investors at this point. And there are a lot of other sources for that, but at the end of the day, the CapEx is the primary driver. And we do a bottom up estimate on the GPUs required, and all of those GPUs are going to require an awful lot of data center capacity and power.

We estimate over the last 30 years almost, we've built out basically close to 60 gigawatts of data center capacity, and that's how we measure data centers, is how much electricity can they basically provide to the chips. We think that'll double in the next three years. So 30 years to build up 60 gigawatts, we're going to need another 60 gigawatts in a three to four year period, which is going to be incredibly complicated to do. It clearly won't be a straight line. We think there's a good chance we run out of capacity in a one to two year period and prices will go up, we think, a lot for data center capacity and for all the components that are in the data centers. And then we think this also drives an increase in overall electric usage on the grid.

But we were lucky enough to partner with a white paper on this subject on AI infrastructure with both Colin and Noah and Rick Schafer, our semiconductor analyst, and basically, our thesis was try to invest in the arm suppliers. And with that, I think Noah follows some of the key component providers inside the data centers where we need more cooling, more power distribution, both inside the data centers and then outside the data centers, which is why I will hand it over to Noah at this point.

Jane Ross:

Yeah, Noah, on that point, if you could take it from there because on this issue of physical infrastructure, we keep hearing about timeframes stretching out to get cooling systems, to get the components, supply

chain issues given demand, and you've spent a lot of time talking about these challenges and potential solutions, so we'd love to hear your thoughts.

Noah Kaye:

Sure. I think at the root of this, we're talking about the challenges of building fiscal infrastructure for AI, and you have increasing compute density driving higher power and thermal management needs, and you have resource bottlenecks including grid infrastructure gating the pace of build out. And so what we've really tried to do is frame up some of those infrastructure challenges for data center owners and the solutions that we're looking for to help solve those challenges. So we think about growing facility level power demand and increasing cost of inefficient power conditioning, sustainability and energy efficiency considerations, which are increasingly of importance for hyperscalers and cool providers. You have the grid infrastructure inadequacy I mentioned earlier. You have an increasing cost of downtime, and then you have server and rack level thermal demands exceeding the limits of air cooling.

Basically, as you get past, call it 20, 25 kilowatts per rack, you need to go to solutions like liquid cooling because of increased thermal management needs. Now, what we've seen from the industry so far is that the arms suppliers, to use Tim's phrase, are generally doing a pretty good job of coordinating, increasingly collaborating with not only the data center developers and operators, but also even the chip makers in terms of charting the technology path, figuring out how can we power and cool the next generation of more powerful chips? And there is, I think, a capacity planning benefit to having these stretched-out time horizons. If we're looking at 18 to 24 months for large projects to be developed, well, that provides some capacity planning windows for the big manufacturers. So, so far, while we definitely see the capacity constraints coming into effect on certain parts of the supply chain, it's not necessarily affected the plays that we cover within the HVAC and power management side.

Jane Ross:

And spend a couple more minutes on that, on the HVAC companies and the whole effort to cool these facilities.

Noah Kaye:

Absolutely. It's easier to conduct heat through liquid than air, and so as you get into rack level cooling capacity, beyond, as I said, 20 kilowatts, you need to go to liquid cooling. There's actually many different types of liquid cooling. There's direct to chip cooling, there's immersion cooling. It's an emerging field. It's quite a small share of the market today, 10% or less, but it could grow to a quarter or more of the market within five years, growing at a 40% CAGR. It is probably the biggest part of the wallet share expansion opportunity per megawatt basis that the companies that we're covering stand to benefit from.

I will say, at the same time, we're not going to transition completely to liquid based cooling for a couple of simple reasons. One, you still need to have heat rejection from the facility, and so that would fall in the role of traditional chillers, air handlers and the like. And two, it's just simply not practical to think that the existing infrastructure is going to undergo a massive reconfiguration of the architecture of these data centers. And so we're really looking at hybrid cooling. That creates opportunities for both traditional players such as the HVAC OEMs—we heard, for example, from Carrier that their data center backlog has doubled in the last couple of month—as well as liquid cooling plays, and companies like Vertiv and Modine are making big investments here to ramp up their capacity significantly and take advantage of the liquid cooling opportunity.

Jane Ross:

And I like that term, hybrid solutions, and Colin, I want to bring it over to you because in terms of the big picture fueling all of these requirements, it isn't going to come from coal, right? How are we going to create the energy for all of this?

Colin Rusch:

Yeah, it's a critical question and a difficult one to answer. As we look at the renewable space and the broader electrification of multiple end markets, whether it's transportation, heating, cooling, which Noah has done a ton of work on, and any number of other applications like into industrial applications where we're seeing electricity and zero emissions solutions come to the fore, there's a major growth of the infrastructure. We're looking at a doubling of the grid, and then also a need to manage more volatility in the grid as you start layering in renewables onto this. So there's a couple of things that we like to remember around our coverage universe. First, power is essentially a real estate play at root, so this is really about where the power is and where it needs to get to and understanding proximity and efficiency around that dynamic.

The second thing is around how do you buffer technology along the way? And this hybrid solution, interim solution is a big one, which is why we focused on energy storage. And what we're seeing on the stationary storage front right now is a tripling of new projects in the last year and incremental growth there. And then beyond that, we're looking at the emergence of virtual power plants with companies like Sunrun, and microgrids with Enphase as a real key supplier of small-scale microgrid functionality. And so the root of this problem is this real estate issue about where you need the power and how you're going to get it there, and I think we're going to see an emergence of newer solutions around grid infrastructure, an incremental drive towards modular solutions that can plug into the grid and island off in part. You're also going to see increased power prices across the network, which some of the best ways to play the infrastructure need on the data centers is really about deliver electricity prices at the edge of the grid in the retail space.

And so residential solar is a great way to get involved here and as a tangential play, and this is a big problem that we don't have a great solution for. One of the more interesting dynamics here is that the interconnection queue for connection to the grid is now extended up to four to five years, where the planning process for a lot of these data centers is three years or fewer, and so there's a dynamic where the grid infrastructure is behind and slower than the real need around data and compute infrastructure needs. And so as we look at this, we just see multi-year growth and replacement cycles driving CapEx in the power sector to levels that are two to three X what they've been historically.

Jane Ross:

And one of the things that the three of you talk about is that given the immensity of all of this, it's not going to happen in a straight line. So I think it'd be helpful for each of you to talk about how you frame the variables or the metrics or the important things that you are watching that are really behind your recommendations and your view ahead. So Tim, can I start with you?

Tim Horan:

Great question. We are going to be focused on how much does it cost to rent out GPUs? It's a balancing act but we do expect cost to decline at a steady pace, but if it drops dramatically, that'll be fairly negative I think for people sticking on how do we monetize AI infrastructure? And then clearly, one of the most important things is just new AI use cases. Microsoft's Copilot in particular is very closely watched right now because they've been very early to the market with AI. We are seeing some major

productivity benefits from Microsoft's Copilot, but we'd like to see good customer uptake because that is kind of expensive. It's like $30 a month to get Copilot, but ultimately, you're going to want to see new consumer killer apps and new consumer killer use cases. We're not sure where they're going to come. I think Colin has some thoughts on that that he can elaborate on.

And then lastly, we really are looking for the forecasts on GPU and CapEx spend. If they remain high, well, then people are going to be extremely optimistic about AI and AI infrastructure, broadly speaking. But then at the end of the day, we are concerned, even if we are hitting the applications and hitting the new use cases that we are going to have bottlenecks in the data centers for a period of time. And what we'll see is prices going up pretty substantially if that is the case, and then you'll see a lot more data center capacity hitting the market probably in very unconventional ways. We saw this before with Bitcoin about six years ago when the price initially went up pretty substantially. We basically ran out of a lot of power in certain locations and data center capacity, but you basically had people building power and data centers wherever they could find it and they ended up meeting the demand. But I'll turn it back to you, Jane.

Jane Ross:

Okay. Well, Colin, Tim referenced you and some thoughts around killer apps. Do you want to take it up next?

Colin Rusch:

Sure. I don't know if it's killer apps, but I think it's killer areas of investment that I would think about.

Jane Ross:

Thank you.

Colin Rusch:

With energy storage. As we look at the Venn diagram of needs coupled with where there's potential innovation, energy storage hits all of these marks. Whether we're looking at large-scale stationary storage on the utility scale side or the need for higher energy density in mobile devices, whether it's phones, computers, or cars, there's a real drive around innovation on the material side within the energy storage space that's seen under-investment on a multi-decade basis. We've really just seen innovation in the storage space come to the fore in the last five to 10 years, and so that's an area where we've tried to carve out some IP from a coverage perspective, but we also see an awful lot of innovation happening and some real winners emerging.

Equinix is one that I mentioned at the start here. They're developing batteries that go into cell phones and computers, and really, they're nearly doubling the energy density on a volumetric basis versus the comparable solutions in many cases when you look at the full package. That's a huge deal for phones. If you're going to be running twice as much compute on those phones, you need that extra power.

On the stationary side, we haven't talked about Tesla, but Canadian Solar has a little bit more direct exposure on this from a corporate perspective in terms of their total growth and stationary storage, but they've been really an aggregator of new technology into stationary needs, and so they're looking at low-cost technologies that are coming out of Asia and bringing them into other geographies to stabilize the grid. And then I will just mention Tesla as an innovator in and around mobile technology for vehicles and mobile energy storage. They've been an innovator around materials for those applications and have begun starting to talk about distributed inference and the capacity for their fleet of their new FSD vehicles to really be part of the latent compute power out in the field, running applications and digesting

information that feeds back into the neural network. And so there's going to be, I think, a lot of fits and starts here, but energy storage and the increased energy density and buffering capacity.

Jane Ross:

Right. It's been a little slow coming, the energy storage side, but Noah, where are you in all this?

Noah Kaye:

Yeah, thanks. Just to pick up on that, Tim mentioned the shipments in GPUs and CPUs, and I think the mix is really important to watch within my coverage, as well as the technology pathway for future generations of chips. Because what we're really talking about here is the power and thermal design power of these chips, because that's going to dictate how the physical infrastructure environment within the data center has to evolve. We talked about before the collaborations that are going on between some of the key suppliers in our coverage, the chip makers and of course the developers on that journey, and so the mix matters very much in terms of the higher compute density environment driving a higher wallet share opportunity for the coverage.

In terms of what that actually looks like for companies, I think it's a fairly straightforward KPI, such as orders, backlog, backlog growth, and book to bill. We just saw Vertiv in their 1Q report 1.5 times book to bill with 60% orders growth, and a big part of that was you had longer cycle orders being placed for this new generation of purpose-built AI data centers, and that longer lead time is feeding into these larger backlogs. So we think those are important indicators to monitor in terms of not only demand but the velocity in which demand comes online, and then how companies are executing capacity expansion to meet that demand.

Jane Ross:

All right. Well, gentlemen, this interdisciplinary group provided a wonderful conversation about some pretty thorny, complicated subjects, so thank you all for your time today.

Tim Horan:

Thanks, Jane.

Noah Kaye:

Thank you, Jane.

Colin Rusch:

Thank you, Jane.